A systematic review of the reliability, construct validity, and responsiveness of health-related quality of life measures within HIV populations

Forest Baker, PhD

Department of Veterans Affairs Health Economic Resource Center

This work was financially supported entirely by a grant from the Department of Veterans Affairs Cooperative Studies Program (CSP # 146) and will be (Oct. 18, 2004) presented at the 26th annual meeting of the Society for Medical Decision Making. Address correspondence to Forest Baker, PhD, Health Economic Resource Center, Department of Veterans Affairs Palo Alto Health Care System, 795 Willow Road (152 MPD), Menlo Park, CA 94025; phone: (650) 493-5000; fax (650) 617-2639; e-mail: forestbaker@hotmail.com.

Conducting a cost-utility analysis requires estimating the health utility for the various health states that are experienced as a result of receiving or not receiving a health intervention. Six different health-related quality of life (HRQL) measures are commonly used to gather empirical estimates of health utilities: the Visual Analog Scale (VAS), Standard Gamble (SG), Time Tradeoff (TTO), Health Utility Index (HUI), EuroQol (EQ-5D), or Quality of Well-Being Scale (QWB). This article articulates criteria for evaluating HRQL measures and the need to establish their psychometric properties within each population of interest. Research that asked HIV-infected populations to assess the desirability of their own health state using at least one of these HRQL measures was reviewed in order to assess the psychometric properties of these measures within HIV populations. Thirty-five articles were included in the review. The majority of the information found in the articles was related to the construct validity of these HRQL measures and there was little to no published research on their reliability or responsiveness. In general, the VAS, EQ-5D, and QWB were highly correlated with health status measures and markers of disease progression, while the SG had weaker correlations with these variables. The TTO and HUI were relatively untested. There is a great need for research on the reliability and responsiveness of these HRQL measures within HIV populations.

KEY WORDS: health-related quality of life, health utilities, reliability, validity, responsiveness, HIV

A systematic review of the reliability, construct validity, and responsiveness of health-related quality of life measures within HIV populations

An attractive feature of a cost-utility analysis (CUA) is that it provides a common metric on which to compare the economic implications of different health interventions. According to guidelines established by a task force of the U.S. Public Health Service, a CUA should score treatment outcomes in terms of the number of quality-adjusted life years (QALYs) that an intervention yields.[1] The number of QALYs associated with an intervention is computed by weighting years of survival by its associated health-related quality of life (HRQL). Estimating survival is relatively objective and straightforward; estimating the HRQL for the health states that a person experiences during those years of survival poses more obstacles. Six different HRQL measures are commonly used to gather empirical estimates of the utility associated with various health states. They are the Visual Analog Scale (VAS), Standard Gamble (SG), Time Tradeoff (TTO), Health Utility Index (HUI), EuroQol (EQ-5D), and Quality of Well-Being Scale (QWB). By reviewing the application of these measures within HIV populations, this article attempts to provide some guidance to researchers who are faced with choosing a HRQL measure to include in their cost-utility study on HIV interventions. Furthermore, this review may serve as a guide for how to assess the psychometric properties of HRQL measures in other populations.

All six HRQL measures involve some procedure that assigns a utility score to a participant’s health state. The VAS consists of one item that asks participants to think about their current health state and then to score it by placing a mark on a vertical line that is anchored between best imaginable health state and worst imaginable health state.

VAS scores can range from 0 to 100, with higher scores reflecting better health states. The SG asks participants to assess how much risk of death they would be willing to take in order to transition from their current health state to perfect health (assuming that they would then live for the rest of their life in their current health state or perfect health).[2] The SG is often administered with a computer or a “chance board” that titrates the risk of death over a number of choice trials in order to assist participants’ judgments. SG scores can range from 0 to 1, with higher scores reflecting better health states (i.e., less risk is taken to escape the current health state). The TTO asks participants to assess how much time they would be willing to subtract from their remaining years of life in order to transition from their current health state to perfect health (assuming that they would then live for the rest of their life in their current state of health or perfect health). The TTO is administered in a similar fashion as the SG and TTO scores can range from 0 to 1, with higher scores reflecting better states (i.e., less time is given up to escape the current health state).[2]

The HUI, EQ-5D, and QWB differ from the VAS, SG, and TTO in that they are multi-item, multi-dimensional measures that, based on participants’ responses, assign participants a specific health state. The health states within the classification scheme are associated with different health utility scores. These utility scores were obtained by asking members of the general public to read descriptions of the different health states and then assign them a health utility score using the VAS, SG, or TTO. Thus, the HUI, EQ-5D and QWB are often referred to as community preference measures since participants’ current HRQL is scored according to the community’s sentiments. Their scores can range from less than 0 (worse than death) to 1, with higher scores reflecting better health states. The HUI (Mark 3) is a 15-item questionnaire that measures 8 domains of health: vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain.[3] The older version of the HUI, the Mark 2, consists of seven domains: sensation, mobility, emotion, cognition, self-care, pain, and fertility. The EQ-5D consists of five items. Each item assesses a different health domain: mobility, self-care, usual activity, pain/discomfort, or anxiety/depression.[4] The QWB is an interviewer-administered questionnaire that has three functioning subscales (mobility, physical activity, and social activity) and additional items concerning symptoms or problems.[5]

Criteria of evaluating HRQL measures

Ideally, any measurement system that scores a subjective construct, such as HRQL, should be reliable, valid, and responsive to changes in the level of the construct within individuals.[6] It is important to note that these psychometric qualities are specific to a measurement's purpose and to the population in which it is administered. For example, a measure that is used to score the mobility of arthritis patients may not be reliable, valid, or responsive to change when it is used for a different purpose (e.g., to measure the desirability of a health state) or administered to a different population (e.g., HIV-infected individuals). Whether these psychometric qualities generalize to other contexts is an empirical question. For example, the estimated reliability of a measure depends directly on the amount of true variability in sample scores, with more true variability allowing for higher reliability estimates. Therefore, these measurement characteristics should be established within each new population of interest rather than assuming that they generalize across different populations. Unfortunately, the latter is common practice within health outcomes research. This review specifically assesses the reliability, validity, and responsiveness to change of HRQL measures within HIV populations.

Reliability

There are two forms of reliability: test-rest and internal consistency. Both are negatively affected by random error, but they are independent in that a measure can have one of these attributes, both, or neither. Test-retest reliability is the stability of a measure's scores across repeated administrations. Test-retest or intra-class correlation coefficients are commonly used to analyze the relationship between repeated administrations of a measure. A correlation of at least .7 is typically considered a minimal reliability criterion.[7] However, high test-retest correlations are not desirable, for example, when participants' true scores have actually changed over time.

The domain-sampling model in psychometrics defines reliability in terms of internal consistency.[8] It assumes that a measure consists of a subset of items that have been randomly sampled from all possible items that make up the domain of interest. Hypothetically, the true score for any individual would be obtained if the measure included all of the possible items within the domain. Internal consistency is the extent to which the total score from a survey correlates with the total "true" score based on all possible items that could be asked about the domain of interest. Chronbach's alpha (α) is used as a measure of internal consistency and an α of a least .7 is typically considered a minimal reliability criterion. Some have argued that α should not be greater than .9 since it suggests that the items in a measure do not cover the full range of the domain.[8]

Reliability is driven by the amount of true variance relative to error variance in the scores.[9] The real concern about using an unreliable measure is that it may not reveal important differences between groups, meaningful changes across time, or significant relationships with other variables. The internal consistency of a measure is likely to improve when more items are added to it or when it is administered to a more heterogeneous population.

Construct Validity

Construct validity examines the logical relationships that should exist between HRQL measures and other variables. This includes monotonic relationships with related variables (convergent validity) and the ability to discriminate between clinically important groups (discriminant validity). For example, if the construct "desirability of a health state" is supposed to be negatively related to the construct "severity of symptoms", then scores on a HRQL measure should be negatively related to measures of symptom severity, such as the number of reported symptoms or the reported severity of disabilities.

Assessing the validity of a measure is more difficult than assessing its reliability. To be valid, a measurement system should produce scores that accurately reflect the construct of interest.[10] Validity requires both empirical evidence and rational argument. That is, rational argument concerning how the desirability of a health state should be measured is needed to guide the search for and interpretation of empirical evidence of validity. A major reason for why there are multiple HRQL measures is because researchers do not agree on how to measure this construct. For example, health states can be defined by a number of domains such as mobility, emotion, social well-being, etc., and it is unclear which domains are necessary for scoring the desirability of a health state. Furthermore, some researchers feel that the judgment concerning the desirability of a health state should incorporate attitudes towards risk (i.e., the SG). There is also much debate over whose sentiments these HRQL estimates should be based upon: patients who are experiencing the health state or the general community, which may not include anyone who has actually experienced the health state. The HUI, EQ-5D, and QWB were all designed to produce health utility estimates that reflect community preferences. A goal of this review was to simply assess the empirical evidence of construct validity, not to compare and contrast theoretical arguments that have been used to justify the validity of different HRQL measures.

The relationships between the six HRQL measures and two health status measures, the Short-Form-36 (SF-36) and the Medical Outcomes Study HIV Health Survey (MOS-HIV), were specifically sought out. The SF-36 and MOS-HIV are multidimensional, generic health status measures (or classification systems), which are widely used in health outcome studies in HIV populations.[11] [12] The 36 items in the SF-36 cover eight sub-domains: physical functioning (ten items), social functioning (two items), role limitations due to physical problems (four items), role limitations due to emotional problems (three items), mental health (five items), energy/vitality (four items), pain (two items), general health perception (five items), and changes in health over the past 12 months (one item). Two summary scales, physical and mental component summary scores (PCS and MCS, respectively), can also be computed. Responses to items are summed and the raw scores are converted to a scale ranging from 0-100, with higher scores reflecting better health.

Two versions of the MOS-HIV have been commonly used. The original version assesses ten domains of health over the previous four weeks and is based on 30 items [13]: general health perception (one item), pain (one item), physical functioning (six items), role functioning (two items), social functioning (one item), mental health (five items), energy/fatigue (four items), cognitive function (four items), health distress (four items), quality of life (one item), and change in health over the previous four weeks (one item). The number of items in the MOS-HIV was later increased to 35 by adding four additional general health perception items and one pain item. [12]. Responses within each domain are summed and then linearly converted to a 0-100 scale, with higher scores reflecting better health. Furthermore, the domain scores can be combined to produce a physical health summary (PHS) and a mental health summary score (MHS).[14] The PHS includes physical function, pain, role function, social function, energy, and general health perceptions; the MHS includes mental health, health distress, quality of life, cognitive function, energy, social function, and general health perceptions.

Responsiveness to change

Responsiveness to change is the ability to detect changes in true scores over time and it can be broken down into two types: external and internal.[15] The *external responsiveness* of a HRQL measure is the degree to which changes in its scores are related to changes in scores on other related variables. For example, one way to assess the external responsiveness of a HRQL measure would be to see whether changes in scores on a measure of mobility are correlated with changes in scores on the HRQL measure. Correlation coefficients are typically used to assess the relationship between change scores.

The *internal responsiveness* of a HRQL measure is the degree to which its scores change after a stimulus that is known to affect a related variable is introduced or removed. For example, one way to assess the internal responsiveness of a HRQL measure would be to give a group of participants a health intervention that has been shown to improve mobility and then see how much HRQL follow-up scores improve from baseline. Internal responsiveness is typically expressed in terms of an effect size (ES, mean change in scores divided by the standard deviation of baseline scores) or a standardized response mean (SRM, mean change in scores divided by the standard deviation of change scores). Internal responsiveness is considered less informative than external responsiveness because internal responsiveness is not specifically related to changes in other related variables. For example, HRQL scores could greatly increase after a treatment is introduced that is intended to improve mobility, however, this increase may not be correlated with changes in mobility.

Methods

Literature search

The literature review was limited to published, peer reviewed studies (e.g., no abstracts or dissertations) that asked people who were infected with HIV to assess the desirability of their current health state using at least one of the following measures: the VAS, SG, TTO, HUI, EQ-5D, or QWB. Furthermore, if a study included both HIV positive and HIV negative participants, it had to specifically present results for the HIV positive participants separately in order to be included in the review. The key words and modifiers "(hiv or aids) and (utility or utilities or health or quality of life)" were combined with an identifying phrase or phrases for each measure (i.e., "rating scale or visual analogue scale or vas", "standard gamble or sg", "time tradeoff or time trade-off or tto", "health utility index or hui", "euroqol or eq-5d", and "quality of well-being scale or qwb") and entered into Ovid and PubMed to search for articles published between 1966 and June, 2004. This search produced 363 citations. The title and abstract of each citation was reviewed and 38 of these citations were selected for a full review. The references in these articles were also searched for relevant papers, which resulted in the inclusion of an additional 8 articles for full review. In the end, a total of 35 articles were included in the review (see Table 1)

The most common version of the VAS asks respondents to evaluate the desirability of their current health state on a 100-point scale that is anchored between worst imaginable health state (0) and best imaginable health state (100).[16] Some studies on HIV populations have used alternative versions of the VAS that, for example, use a 10-point scale or that transform the sum of several items into a 100-point scale.[17-20] This review included only versions of the VAS that asked respondents to evaluate their current overall health or well-being on a 100-point scale. This constraint was implemented in order to avoid unnecessarily introducing variability into the findings due to differences in presentation format. Even with this constraint, there was more published information on the VAS than for any of the other HRQL measures.

Lastly, the literature search revealed that some data were published in more than one paper. [14, 21-24] Bult et al. (1988) tested different regression equations on the data reported by Tsevat et al. (1996), but did not report any additional findings that were relevant to this review, so it was not included in the review. Lenert and colleagues published three papers on the same set of data: one reported on the feasibility of using a computer to present the HRQL measures and the other two reported on the reliability and validity of these measures. Therefore, only the latter two studies were included in the review. Two papers by Revicki and colleagues (1995 and 1999b) each reported on two separate studies. A third paper by Revicki and colleagues (1998), however, presented additional analyses on both of these studies using slightly different sample sizes. This third paper was included in the review, but it was not viewed as a separate study. Therefore, this review was based on a total of 35 articles and 33 unique studies/data sets.

Evidence of reliability

Articles were searched for information on the reliability of the six HRQL measures. This included estimates of internal consistency, as measured by Cronbach's alpha, alternative forms correlations, and test-retest reliability, as measured by Pearson or intra-class correlation coefficients. A correlation of at least .7 was used as the minimal cutoff for reliability.

Evidence of construct validly

Evidence of construct validity was guided by three assumptions: 1) there should be a positive relationship between HQRL measures, 2) HRQL measure scores should be positively related to health status measure scores, and 3) HRQL measures should be related to clinical status classification, number of experienced symptoms, and other clinical markers of disease progression. Articles were searched for any reported, relevant relationships between any of the HRQL measures and other variables. Correlation coefficients (Pearson's r and Spearman's rho) were taken directly from published results or an estimate of Pearson's r was computed if sufficient information was provided. When effect size was computed in terms of r, .10 was considered small, .30 was considered medium, and .50 was considered large.[27]

The correlation between two measures is constrained by the internal-consistency of each measure (as measured by Cronbach's α). For example, if two constructs share a perfect linear relationship (*r* = 1.0), but the reliability of the measures for each construct is .7, the obtained correlation between these measures cannot be higher than .7. Furthermore, with this level of reliability, a large effect size between these two variables (*r* = .5) would appear as only a medium effect (*r* = .35) Thus, in order to understand how much this relationship is attenuated, it is important know the estimated reliabilities for two measures when assessing the strength of the relationship between them. Given that this review was interested in the correlations between the HRQL measures and the SF-36 and MOS-HIV, articles were searched for reported Cronbach's α for the SF-36 and the MOS-HIV.

Evidence of responsiveness

Articles were searched for evidence of internal and external responsiveness to change for each of the six HRQL measures. For external responsiveness, Pearson's *r* was taken directly from published results or an estimate was computed if sufficient information was provided. For internal responsiveness, ES's and SRM's were taken directly from articles or were computed if sufficient information was provided. For ES and SRM, .2, .5, and .8 were considered benchmarks for small, medium, and large effects, respectively. [27]

Results

Table 1 provides information on the 36 articles included in the review. Across all of the studies, participants were predominately male and in their 30's. Unless stated otherwise, all participants in each study were HIV-positive. For each study, it is informative to note the year that it was published, its sample size, and the severity of HIV in its sample. Because of the availability of anti-retroviral medications, HIV symptoms are better managed today than in the past.[28] This suggests that, compared to more recent studies, older studies on HIV patients may be more likely to find lower HRQL scores and larger differences in HRQL scores between people who are aymptomatic and symptomatic. Also, a study with a more heterogeneous sample is more likely to produce larger effect sizes than less heterogeneous samples.

Few studies stated whether the HIV classification schema was based on the US Center for Disease Control's (CDC's) 1986 or 1993 classification system. It is probably safe to assume that studies published several years after 1993 used the newer guidelines, while earlier studies clearly used the older guidelines. An important difference between these guidelines is that the 1986 classification system is entirely symptom based, while the 1993 system also includes CD4 cell count, so that people with CD4 cell counts below 200 are also classified as having AIDS regardless of their symptoms. Within both classification systems, if a person transitions to a symptomatic or AIDS classification, they can never return to a less severe classification even if their symptoms abate or if their CD4 cell count increases. Thus, the CDC HIV classification systems should not be viewed as perfect indicators of current disease progression or health status.

The six HRQL measures were not used with equal frequency. The VAS was used in almost twice as many studies (n = 18) compared to second most common measure, the QWB (n = 10). The frequency of use for the rest of measures in descending order was as follows: EQ-5D (n = 8), SG (n = 5), TTO (n = 4), and HUI (n = 2).

Reported evidence of reliability

There is virtually no information on the reliability of the six HRQL measures within HIV populations. None of the articles in the review reported the internal consistency for any of the three multi-item HRQL measures (EQ-5D, HUI, or QWB). QWB score involve some judgment on the part of the interviewer, but none of the articles reported inter-rater reliability or, for that matter, even mentioned how the study design took this into account. For test-retest reliability, only one article reported modest intraclass correlation coefficients for the VAS (.55), TTO (.42), and SG (.47).[23] This study, however, was not specifically designed to test the stability of these scores. It was based on a medium-sized convenience sample (*n* = 75) of HIV clinic patients. Some were reported to have less than “relatively well-controlled HIV-infection”, the retest took place 3 to 6 weeks later, and patients were not asked whether they felt that their health had changed since the baseline test, which is important for interpreting test-retest reliability estimates.

Reported evidence of construct validly

The amount of information pertaining to the construct validity for each of the six HRQL measures was strongly related to the number of studies that used the measure. Table 2 shows the obtained correlations between the six HRQL measures. The VAS appears to have large correlations with the TTO, EQ-5D, and QWB. (In addition, another study reported a significant correlation of .26 between a VAS that was scored from 0 to 10 and the SG).[29] The TTO appears to have a medium correlation relationship with the SG and QWB. Without reliability estimates for the HRQL measures, however, it is difficult to know the degree to which these correlations are attenuated and whether the true size of these correlations is higher than those reported. There was no information on the HUI.

The obtained correlations between the six HRQL measures and the SF-36 and MOS-HIV are shown in Tables 3 and 4, respectively. Cronbach's α was quite high for both the SF-36 and the MOS-HIV. In addition, another study reported that Cronbach's alpha ranged from .84 to .93 for the total score and the eight sub-domains of the SF-36.[30]. For the multi-item, sub-domains of the MOS-HIV, Revicki, Wu, and Murray (1995) and Copfer, Ampel, and Hughes, et al. (1996) also reported that Cronbach's alpha ranged from .77 to .94 and .85 to .95, respectively.[22, 31]. Thus, low correlations between these two measures and the HRQL measures are probably due to low reliability in the HRQL measures or weak linear relationships between these variables.

The VAS and EQ-5D had large correlations with each of the sub-domains of the SF-36, except for with Bodily Pain, which had a medium correlation with both health utility measures (see Table 3). The QWB also had large correlations with each of the sub-domains of the SF-36, except that the correlation with Role Limit Emotional was moderate and the two reported correlations with Mental Health greatly differed (.04 and .41). The size of the correlations between the SF-36 sub-domains and the TTO ranged from small to large. The correlations between the SF-36 sub-domains and the SG were the smallest of all the measures, ranging from small to medium. No correlations were reported between the HUI and the SF-36.

As with the SF-36, the VAS, EQ-5D, and QWB had medium to large correlations with each of the sub-domains of the MOS-HIV (see Table 4). The VAS and EQ-5D also had large correlations with the physical and mental summary scores. The HUI2 also exhibited medium to large correlations with the MOS-HIV sub-domains. One study reported a small correlation between the SG and the General Health Perception subdomain and that the rest of the correlations were not significant. Given that the sample size of this study was 123 and that at a $t > 1.984$ is needed for α to be $< .05$, two-tailed test, the obtained correlations with the SG all had to be less than .19. No correlations were reported between the TTO and the MOS-HIV.

A number of studies reported the relationship between scores on the EQ-5D and the Karnofsky Performance Scale (KPS), a functional status measure that focuses on physical performance and dependency.[32] The reported correlations between the EQ-5D and the KPS ranged from medium to large (.22[31], .41[33], .44[34], and .51[35]).

Table 5 shows the reported effect sizes of a number of HIV-related clinical diagnoses on the six HRQL measures. The effect of the CDC's HIV classification (i.e., asymptomatic, symptomatic, and AIDS) on VAS scores was inconsistent across studies. For example, the effect size between asympomatic and symptomatic patients ranged from very small to very large across three studies. Procedural differences across these studies may have been responsible for the inconsistent findings. Participant's in Tsevat et al.'s (1996) study were interviewed over the phone and the VAS was anchored at death to excellent health; during a structured in-person interview, participants in Revecki et al.'s (1995) study first rated several hypothetical HIV health states before rating their own health state on a VAS that ranged from death to complete health; and participants in Schag et al.'s (1992) study self-completed a questionnaire in which the VAS asked participants to score their quality of life on a scale that ranged from very low to very high. Thus, it is difficult to determine how well the VAS discriminates between the CDC HIV classifications. There seems to be a small to medium effect of HIV classification on SG scores and only a small effect on TTO scores, however, these finding have not been replicated across multiple studies. Furthermore, SG scores were unexpectedly higher for symptomatic patients than for asymptomatic patients. Based on one study, HIV classification reportedly has medium to large effects on HUI2 scores. The effect size of HIV classification on EQ-5D scores could not be calculated from the papers included in the review except for a medium effect size between HIV- and HIV+ participants. HIV classification generally has a medium effect on QWB scores.

Collapsing across the various health utility measures, it appears that health utility is lower for AIDS patients compared to cancer patients, but having cancer in addition to AIDS has little effect on health utility. Within hemophiliacs, being HIV+ seems to have a medium to large negative effect on health utility regardless of the severity of hemophilia. The effect of diagnosis of Cytomegalovirus Retinitis on health utility within AIDS patients was mixed, ranging from almost no effect to a medium effect.

Table 6 shows the reported or estimated effect size of common markers of HIV progression on the six health utility measures. There is a consistent small to medium effect of CD4 cell count on VAS scores. One study reported a medium to large effect on TTO scores. The effect size of CD4 cell count on the SG, EQ-5D, and QWB was inconsistent across studies, with one study reporting a small effect and the other(s) reporting a medium effect. One reason for this difference may be due to a time lag from when the CD4 cell count was taken to when the health utility measure was completed. Less time between the two measures may have resulted in stronger associations. For example, in Copfer et al.'s (1996) study the QWB was administered up to 3 months after the CD4 cell count was taken and there was only a .04 correlation between these two variables. Collapsing across the health utility measures, it appears that viral load and serum Beta-2 counts have small to medium effects on health utility. No studies reported on the effect of biological markers on HUI scores.

Reported evidence of responsiveness to change

Table 7 shows reported evidence of external and internal responsiveness for the HRQL measures. Only the VAS and EQ-5D had any published information concerning external responsiveness. Change in VAS scores was weakly related to change in CD4 cell count, viral load, and presence of adverse symptoms. The correlation between VAS change scores and CD4 cell count change scores differs from the correlations between these two variables that are presented in Table 6 – those correlations are based upon static (one point in time) VAS and CD4 cell count scores. Change in EQ-5D scores had a small to medium relationship with change in the presence of adverse symptoms. There was scant published information on the internal responsiveness of the HRQL measures, except for the VAS. The VAS showed medium changes across time for people who had developed an opportunistic infection or had AIDS. VAS scores showed only a small change across time within asymptomatic or symptomatic patients. Studies that administered medication to patients generally found that VAS scores showed small decreases across time. SG scores generally showed small decreases across time within disease classification. EQ-5D scores showed small decreases for people who had developed and opportunistic infection. Surviving patients showed a small to medium increase in their QWB after taking Zidovudine for an average of 19 weeks, while patients receiving placebo showed almost no change. No studies reported on the responsiveness of the TTO or the HUI.

Three studies reported change in health scores, but not in terms of correlations, effect sizes, or standardized response means. Revicki et al. (1995) found that for each additional HIV-related symptom that developed between baseline and a 4-month follow-up, VAS scores significantly deceased by 7.5 and SG scores decreased (but not significantly) by 4.67 after adjusting for demographic variables and differences in baseline clinical variables and baseline health status scores. Another study, based on seven HIV-positive participants who transitioned to a worse disease stage, found that the mean percent change in the VAS, QWB, and TTO was –9.3, -5.1, and 2.0, respectively; suggesting that, on average, TTO scores failed to show a decrement in health utility.[26] Carr et al.'s (2000) study found that EQ-5D scores significantly increased in three groups of patients receiving different medications.

Discussion

A number of HRQL measures are available to researchers who wish to gather empirical estimates of health utilities for CUA's. Researchers should select measures that are reliable, valid, and responsive to change within the population that they intend to study. This review presented all of the published evidence for the reliability, validity, and responsiveness to change of six HRQL measures within HIV populations. Unfortunately, this review highlighted the fact that, while there is information on the construct validity for some of these measures, there is little to no evidence available on the reliability or responsiveness to change for most of these measures within HIV populations. This greatly constrains the ability to make recommendations for the use of one or more HRQL measures in lieu of the others. Until this evidence is available, HIV outcomes researchers should at least be aware of the scant psychometric information that is available on these measures.

Visual Analog Scale

There is far more information available on the psychometric properties of VAS within HIV populations than any of the other five HRQL measures. There is strong evidence for the construct validity of the VAS. The VAS correlates strongly with the TTO, EQ-5D and QWB; however, except with the EQ-5D, these results have not been replicated. Seven different studies have found medium to large correlations between the VAS and the SF-36 or MOS-HIV (including their summary scores and subscales). Three studies have shown that biological markers of disease progression consistently have only a small to medium effect on VAS scores. The effect of HIV classification on VAS scores is more varied, ranging from small to large effects. These disparate findings highlight the possibility that VAS scores are sensitive to presentation format, which would be avoided if researchers used a common format, such as the VAS that is included in the EQ-5D. The VAS appears to be only weakly to moderately responsive to changes in health and the reliability of the VAS is basically untested. Only one study reported a modest test-retest correlation. The psychometric properties of the VAS appear to be at least as good, if not better, than those of the other two single-item HRQL measures (i.e., the SG and TTO).

Standard Gamble

There is little information available concerning the psychometric properties of the SG within HIV populations. The reliability of the SG is basically untested, with one study finding only a modest test-retest correlation. It appears that the SG has only weak evidence of construct validity (according the three criteria discussed in the introduction).

SG scores have a medium correlation with TTO scores, but SG scores appear to have only small to medium correlations with SF-36 and MOS-HIV scores. HIV classification and CD4 cell count have small to medium effects on SG scores, but more research is needed since one of the studies found that SG scores were unexpectedly lower in asymptomatic patients than in symptomatic patients. One study reported internal responsiveness to change for the SG and found that it showed small changes across time.

Time Tradeoff

There is also little information available concerning the psychometric properties of the TTO within HIV populations. As with the VAS and SG, the reliability of the TTO is basically untested, with one study finding only a modest test-retest correlation. There was, however, some evidence concerning the construct validity of the TTO. It has medium to strong correlations with the SG, VAS, and QWB. One study found that the TTO has small to large correlations with the SF-36 subscales, two studies found that HIV classification has only a small effect on TTO scores, and one study found that CD4 cell count has a medium effect on TTO scores. The TTO's responsiveness to change is untested.

EuroQol

Several studies suggest that the EQ-5D has construct validity in that it consistently correlates strongly with the VAS and the SF-36 and MOS-HIV subscales. Being HIV-positive has a medium to large negative effect on EQ-5D scores and biological markers of disease progression seem to have small to medium effects on EQ-5D scores. One study suggests that increases in symptoms are reflected by small to medium changes in EQ-5D scores. Thus, the psychometric properties of the 5-item EQ-

5D appear to be at least as good, if not better, than those of the other longer multi-item HRQL measures (i.e., the HUI and QWB). There is no information concerning the internal consistency or test-retest reliability of the EQ-5D within HIV populations.

Health Utility Index

Compared to the other multi-item HRQL measures, the HUI has the least amount of information concerning it psychometric properties within HIV populations. There is no information concerning the internal consistency, test-retest reliability, or responsiveness to change of the HUI. Only one study reported that the Mark 2 version of the HUI correlated strongly with the MOS-HIV subscales and that CDC classification generally had medium to large effects on HUI2 scores.

Quality of Well-Being

Several studies suggest that the QWB has construct validity. It has medium to strong correlations with the VAS and TTO and it consistently correlates strongly with the SF-36 and MOS-HIV subscales. HIV infection status and CDC classification have medium to large effects on QWB scores and biological markers of disease progression seem to have small to medium effects on QWB scores. There is no evidence of the QWB's external responsiveness, but one study suggests that patients receiving active medication appear to have medium to large improvements in their QWB scores across time. Thus, while the QWB's responsiveness to change and reliability is relatively untested, there is a fair amount of positive evidence for the construct validity of the QWB. There is no information concerning the internal consistency or test-retest reliability of the QWB within HIV populations.

Conclusions

This review articulated criteria that are commonly used to evaluate measurement systems and the need to apply them within each new population of interest. The research literature suggests that the psychometric properties of HRQL measures are typically assessed within a particular population, but then, they are administered to other populations in which the reliability, validity, and responsiveness of these measures have not been established. In fact, over 30 studies have administered HRQL measures to HIV populations and interpreted the findings with little to no information on the reliability or responsiveness of these measures within this population. This is particularly alarming in studies that are trying to gather empirical estimates of health utilities for CUA's. Using a HRQL measure that not reliable, valid, or responsive to change may result in type I or type II errors and produce inaccurate QALY estimates.

More research is needed on the psychometric properties (especially concerning reliability and responsiveness to change) of the VAS, SG, TTO, EQ-5D, HUI, and QWB within HIV populations in order to determine whether some measures are superior to others. At this point, among the three single-item HRQL measures, the VAS has been tested the most and there appears to be strong evidence of construct validity. Among the three multi-item HRQL measures, the EQ-5D and QWB have been used more frequently and the results suggest that they may also have desirable psychometric properties. While the SG has not been used very often, its preliminary results raise questions about its construct validity. These tempered conclusions, however, are based on relatively limited research and very simple theoretical assumptions.

This review did not scrutinize the theoretical underpinnings of the six HRQL measures, which are necessary for establishing the validity of a measure. Whether a HRQL measure appears to possess desirable psychometric properties will depend on the theoretical approach that one uses to measure health utilities. The goal of this review was to provide all of the findings that are relevant to analyzing the psychometric properties of commonly used HRQL measures, regardless of which theoretical approach is thought to be the most appropriate for measuring health utilities. The empirical evidence presented in this review, however, may be used to assess these theoretical assumptions.

554

Table 1

Articles Included in the Review

Study #	First author	Year	*n* (% Male)	Age Mean (Variability)	CD4 count (cells/mm^3) Mean (Variability)	Description and CDC classification of HIV participants[a]	HRQL measures in study
[36]	Anderson	1998	HIV+: 99 Cancer: 74 Serious Disease: 28 Total: 201 (78)	HIV+: 37 (24-65) HIV-: 64.6 (22-91)		Asym = 0% Sym = 0% AIDS = 100%	QWB
[37]	Badia	1999	HIV+: 558 (75) HIV-: 80 (83)	HIV+: 36 (*SD* 7.4) HIV-: 36.3 (*SD* 12.7)	200<Median<500	Asym = 37% Sym = 32% AIDS = 31%	VAS EQ-5D
[38]	Barr	2002	All hemophiliacs HIV+: 23 HIV-: 75	33.9 (13-87)		HIV+ were all severe hemophiliacs with hepatitis B or C	HUI
[39]	Bayoumi	1999	73 (92)	41.35 (*SD* 8.4)	239.4 (*SD* 208.9)	Asym = 12%, Sym = 35% AIDS = 53%	HUI

[40]	Carr	2000[b]	106 (93)	38 (*SD* 9.7)	285 (*SD* 184)	AIDS = 17%	EQ-5D
[31]	Copfer	1996	65 (98)	41[c] (*IQR* 36-46)	180[c] (*IQR* 88-387)	Asym = 42% Sym = 29% AIDS = 29%	QWB
[41]	Delate	2001	242 (85.5)	39.8 (*SD* 8.4)	343[c] (1 – 1826)		VAS EQ-5D
[42]	Djulbegovic	1996	All hemophiliacs HIV+: 18 (100) HIV-: 18 (100)	HIV+: 31.4 (*SD* 8.7) HIV-: 30.5 (*SD* 13.8)	292 (*SD* 220)	Higher frequency of severe hemophilia in HIV+ than in HIV-	QWB
[33]	Hughes	1997	100 (100)	35.3 (23-56)	403 (8-1,016)	Asym = 42%[d]	QWB
[43]	Johnson	1996	80 (99)	30<Median<39	All had >200	All were either Asym or Sym (AIDS = 0%)	TTO
[5]	Kaplan	1989[b]	31 (87)	35.8[c]	153.2 (*SD* 160.2)	AIDS = 42%[a]	QWB
[35]	Kaplan	1995	HIV+: 400 (100) HIV-: 114 (100)			Asym = 68% Sym = 20% AIDS = 12%	QWB
[44]	Kempen	2003	971 (83)	42	<u>CMV Retinitis</u>: None: 163.5 Long-standing: 166 New Diagnosis: 24	AIDS = 100%	VAS EQ-5D

[29]	Lamping	1994[b]	81 (93)	36 (25-64)	All had < 500	Asym = 17% Sym = 36% AIDS = 28% Missing = 9%	SG
[24] [23]	Lenert Lenert	2002 2002	75 (96)	40<Median<50	<300: *n* = 27 >300: *n* = 41 Unknown: *n* = 7		VAS TTO SG
[30]	Miners	1999	All hemophiliacs HIV+: 31 HIV-: 135	58.4 (*SD* 16.1)		HIV+: 27 had severe and 4 had mild or moderate hemophilia	VAS EQ-5D
[28]	Miners	2001	149 (89)	36 (*SD* 9)	353[c] (0-1,991)	AIDS = 30%	VAS EQ-5D
[45]	Mrus	2003	299				VAS SG
[46]	Nickel	1996[b]	57 (93)	30>Median<39			QWB
[47]	Nordic Med. Research Council	1992[b]	474 (91)	38[c]		Asym = 31% Sym = 52% AIDS = 27%	VAS
[48]	Olivia	2003	32 (53)	37.7[c] (*SD* 5.4)			EQ-5D
[22]	Revicki	1995	160 (66) Results based on: 123	36.7 (*SD* 8.3)	276.6 (*SD* 181.5)	Asym = 25% Sym = 31% AIDS = 44%	VAS SG

[14]	Revicki Overlaps with 1995 & 1999b	1998[e]	1995: 162 (66) - Results based on: 154 1999b: 1,022 (93)	1995: 36.7 1999b: 38.7 (*SD* = 8.4)		1995: Asym = 25% Sym = 31% AIDS = 44%	1995: VAS 1999b: VAS
[49]	Revicki	1999[b]	993 (83)	36.5 (*SD* 9)	199.6[c] (*SD* 91)		VAS
[21]	Revicki	1999[b]	940 (91.9)	38.6 (*SD* 8.4)	170 (*SD* 85)		VAS
[50]	Schag	1992	318 (96)	38 (22-68)		Asym = 37% Sym = 20% AIDS = 25% AIDS + Cancer = 18%	VAS
[51]	Tramarin	1992[b]	42 (78.6)	28<Median<33 (*SD* 6.87)			QWB
[52]	Trippoli	2001	All hemophiliacs HIV+: 12 HIV-: 44	38.7 (*SD* 15.4)		Hemophilia: Mild/Moderate: $n = 1$ Severe: $n = 11$	VAS EQ-5D
[26]	Tsevat	1996	139 (94.2)	34.6 (*SD* 6.5)	335.2 (*SD* 282.8)	Asym = 33.8% Sym = 28.1% AIDS = 38.1%	VAS TTO QWB
[53]	Tsevat	1999	51 (71)	36[c] (24-67)	185[c] (*IQR* 106-431)	Asym = 31.4% Sym = 3.9% AIDS = 64.7%	VAS TTO SG

[34]	Wu	1990[b]	32 (87.5)	35.8	152 (*SD* 159.1)	All had AIDS or severe AIDS related complex	QWB
[54]	Wu	1997[b]	68 (91)	39.6 (*SD* 7.2)	All had < 100	All were stable with advanced HIV	VAS
[55]	Wu	2002[b]	990 (94)	38.5 (*SD* 7.8)	All had < 100		VAS EQ-5D
[56]	Zinkernagel	2001	318 (66)	38[c] (20-76)	445 (0-1529)	Asym = 37% Sym = 37% AIDS = 26%	VAS

CDC = US Center for Disease Control; HRQL = health-related quality of life; HIV+ = HIV-positive; HIV- = HIV-negative; Aysm = asymptomatic; Sym = symptomatic; VAS = Visual Analogue Scale; TTO = Time-Tradeoff; SG = Standard Gamble; EQ-5D = EuroQol; QWB = Quality of Well-Being; HUI = Health Utility Index; IQR = Inter-Quartile Range; SD = Standard Deviation

Note. Unless stated otherwise, all participants in each study were HIV-positive.

[a]HIV classification scheme is based on either the CDC's 1986 or 1993 classification system.

[b]Results are based on clinical trial data.

[c]Median

[d]Converted from 1986 CDC classification

[e]See text in Methods section

Table 2

Correlations Between HRQL Measures

	VAS	SG	TTO
VAS	----	----	----
SG		----	----
TTO	.51 [26]	.37 [53]	----
EQ-5D	.63 [41] .59 [28]		
QWB	.44 [26]		.43 [26]

Note. The superscript indicates the study. VAS = Visual Analogue Scale; SG = Standard Gamble; TTO = Time Tradeoff; EQ-5D = EuroQol; QWB = Quality of Well-Being Scale.

Table 3

Internal-Consistency of SF-36 and Correlations with HQRL Measures

	Cronbach's Alpha	VAS	SG	TTO	EQ-5D	QWB
SF-36 Total	.92 [29]		.35 [29]			
Physical Sum Score		.51 [52]			.59 [52]	
Mental Sum Score		.63 [52]			.59 [52]	
General Health Perception	.78 [29]	.71 [52] .66 [26]	.21 [29]	.58 [26]	.76 [52]	.64 [42] .59 [26]
Physical Function	.88 [29]	.61 [52] .51 [26]	.26 [29]	.51 [26]	.64 [52]	.67 [26] .38 [42]
Social Function	.69 [29]	.64 [52] .49 [26]	.25 [29]	.46 [26]	.65 [52]	.63 [26] .61 [42]
Role Limit Physical	.82 [29]	.57 [52] .47 [26]	.19 [29]	.53 [26]	.57 [52]	.65 [26] .30 [42]
Bodily Pain	.85 [29]	.34 [26] .31 [52]	.28 [29]	.34 [26]	.43 [52]	.67 [26] -.41 [42]
Vitality / Energy	.87 [29]	.65 [52] .50 [26]	.40 [29]	.51 [26]	.61 [52]	.68 [26]
Mental Health	.76 [29]	.65 [52] .45 [26]	.23 [29]	.23 [26]	.65 [52]	.41 [26] .04 [42]
Role Limit Emotional	.71 [29]	.58 [52] .33 [26]	.18 [29]	.06 [26]	.55 [52]	.26 [26]

Note. The superscript indicates the study. VAS = Visual Analogue Scale; SG = Standard Gamble; TTO = Time Tradeoff; EQ-5D = EuroQol; QWB = Quality of Well-Being Scale.

558

Table 4

Internal-Consistency of MOS –HIV and Correlations with HRQL Measures

	Cronbach's Alpha	VAS	SG	HUI2	EQ-5D	QWB
Physical Score		.72 [28]a .63 [41]a .59-.57 [55]a .56 [22] ([14]b) .56 [21] ([14]b)			.68 [41] .61 [55]a .59-.72 [28]a	
Mental Score		.63 [21] ([14]b) .60 [55]a .54 [22] ([14]b) .46 [41]a .59-.72 [28]a			.58 [55]a .44 [41] .59-.72 [28]a	
General Health Perception	.88 [28]a .86 [41]a .80 [37]e .57 [33]c	.66 [55]a .66 [37]e .60 [22]d	.14 [22]d	.50 [39]a	.50 [55]a	.57 [33]c .54 [31]c
Physical Function	.92 [28]a .89 [41]a .85 [31]c .83 [37]e .58 [33]c	.45 [55]a .34-.60 [22]d		.58 [39]a	.50 [55]a	.66 [31]c .62 [33]c
Social Function	N/A	.55 [22]d .50 [55]a		.56 [39]a	.49 [55]a	.65 [31]c .38 [33]c
Role Function	.89 [33]c .83 [37]e .80 [41]a .80 [28]a	.39 [55]a .34-.60 [22]d		.37 [39]a	.45 [55]a	.67 [31]c .64 [33]c
Cognitive Function	.92 [41]a .92 [28]a .84 [37]e .68 [33]c	.33 [55]a .34-.60 [22]d		.55 [39]a	.40 [55]a	.59 [31]c .44 [33]c
Pain	.87 [28]a .84 [41]a .78 [37]e	.44 [55]a .34-.60 [22]d		.65 [39]a	.63 [55]a	.52 [31]c .42 [33]c

Energy	.90 [28]a .84 [41]a .79 [37]e .62 [33]c	.60 [55]a .34-.60 [22]d		.64 [39]a	.55 [55]a	.67 [31]c .47 [33]c
Mental Health	.95 [31]c .88 [41]a .87 [37]e .86 [33]c .84 [28]a	.46 [55]a .34-.60 [22]d		.45 [39]a	.50 [55]a	.53 [31]c .23 [33]c
Health Distress	.94 [28]a .91 [41]a .89 [37]e .63 [33]c	.46 [55]a .34-.60 [22]d		.35 [39]a	.48 [55]a	.47 [31]c .24 [33]c
Quality of Life	N/A	.49 [55]a .34-.60 [22]d		.61 [39]a	.48 [55]a	.58 [31]c .37 [33]c

Note. The superscript indicates the study. VAS = Visual Analogue Scale; SG = Standard Gamble; HUI = Health Utility Index; EQ-5D = EuroQol; QWB = Quality of Well-Being Scale.

[a]Used 35-item MOS-HIV [12]

[b]Used 30-item MOS-HIV [13]

[c]Used the 30-item MOS-HIV, but added four general health perception items.

[d]Used the 35-item MOS-HIV, but added one cognitive function item and one energy item.

[e]Used the 35-item MOS-HIV, but assessed health over the previous 2 weeks instead of 4.

Table 5

Effect Size (r) of Diagnoses on HRQL Measure Scores

Comparisons	VAS	SG	TTO	HUI2 / HUI3	EQ-5D	QWB
HIV+ vs. HIV-	.76[28] .24[56]				.41[28]	.30[35]
HIV+ vs. AIDS	.29[53] .20[21]	.24[53]	.10[53]	.40[39] / -----		.42[35]
Asym vs. Sym	.93[22] .21[50] .03[26]	-.26[22]	.16[26]	.29[39]/ -----		.28[26] .26[35]
Asym vs. AIDS	.94[22] .31[50] .12[26]	.14[22]	.13[26]	.49[39] / -----		.41[35] .39[26] .28[31]
Sym vs. AIDS	.42[50] .08[26] -.29[22]	.42[22]	-.03[26]	.27[39] / -----		.43[35] .14[26]
AIDS vs Cancer						-.20[36]
AIDS vs AIDS with cancer	-.003[50]					
SCSAH for AIDS Stage 2 vs stage 3						.43[51]
Mild to severe Hemophiliacs HIV- vs. HIV+	.39[52]				.32[52]	
Severe Hemophiliacs HIV- vs. HIV+				.31[38] / .24[38]		
AIDS & CMV status No vs. Long-standing	.03[44]a				-.03[44]	

AIDS & CMV status Long-standing vs. Newly diagnosed	.19[44]	-.03[44]

Note. Effect size estimates were based on information provided in article. The superscript indicates the study. SCSAH = Severity Classification System for AIDS Hospitalization; [57] CMV = Cytomegalovirus Retinitis; VAS = visual analogue scale; TTO = time-tradeoff; SG = standard gamble; HUI2 = Health Utility Index Mark 2; HUI3 = Health Utility Index Mark 3; EQ-5D = Euroqol; QWB = Quality of Well-Being Scale; Asym = asymptomatic; Sym = symptomatic.

Table 6

Effect Size (r) of Biological Markers of Disease Progression on HRQL Measure Scores

Biological Marker	VAS	SG	TTO	EQ-5D	QWB
CD4 Cell Count	.18[41] .18[56] .18[24]	.25[24] -.01[29]	.37[24]	.40[48] .12[41]	.42[35] .41[33] .04[31]
HIV-1 RNA Level	-.19[41]			-.13[41] >-.34[48]	
Serum Beta-2 Count					-.40[33] -.18[35]

Note. Effect size estimates were based on information provided in article. The superscript indicates the study. VAS = visual analogue scale; TTO = time-tradeoff; SG = standard gamble; EQ-5D = Euroqol; QWB = Quality of Well-Being Scale.

566

Table 7

External and Interval Responsiveness of HRQL Measures

	Change in:			
	VAS	SG	EQ-5D	QWB
External responsiveness				
Increase in CD4 cell count over 24 weeks	.16[21]r			
Increase in log RNA over 24 weeks	-.10[21]r			
Increase in the number of adverse experiences of grade 2 or higher (e.g., neutropenia, anemia, diarrhea, nausea and vomiting, pneumocystis pneumonia, Kaposi's sarcoma) over 4 weeks	-.15[55]r		-.20[55]r	
Internal responsiveness				
After 4 weeks within participants who developed an opportunistic infection	-.50[55]d -.63[55]s		-.33[55]d -.20[55]s	
After 4 months within participants with a certain disease classification:				
Asymptomatic	-.12[22]d -.06[22]s	-.08 [22]d -.04 [22]s		

Symptomatic	-.12[22]d -.06[22]s	-.40[22]d -.20[22]s	
AIDS	-.76[22]d -.38[22]s	-.16[22]d -.08[22]s	
After 48 weeks within participants receiving one of the following treatments:			
Zalcitabine	-.34[21]d* -.34[21]s*		
Saquinavir	-.32[21]d* -.32[21]s*		
Zalcitabine/saquinavir	-.16[21]d* -.16[21]s*		
After an average of 19 weeks within participants receiving on of the following treatments:			
Placebo			.02[34]d
Zidovudine			.40[34]d
After 48 weeks within participants receiving one of the following treatments:			
Zalcitabine/Zidovudine	-.14[49]d -.15[49]s		
Saquinavir/Zidovudine	-.15[49]d -.15[49]s		

Saquinavir/Zalcitabine/Zidovudine	.07[49]d .06[49]s

Note. Effect size estimates were based on information provided in articles. The superscript indicates the study. VAS = visual analogue scale; SG = standard gamble; EQ-5D = Euroqol; QWB = Quality of Well-Being Scale.

[d] *d* = Cohen's measure of effect size: $(M_2 - M_1)/SD_1$.

[s] SRM = Standardized Response Mean: $(M_2 - M_1)/SD_{2-1}$.

[*] Means were adjusted for baseline HRQL score, treatment group, region, and CD4 strata.

References

1. Gold, M.R., et al., *Cost-Effectiveness in Health and Medicine.* 1996.
2. Torrance, G.W., *Utility approach to measuring health-related quality of life.* J Chronic Dis, 1987. **40**(6): p. 593-603.
3. Torrance, G.W., et al., *Multi-attribute preference functions. Health Utilities Index.* Pharmacoeconomics, 1995. **7**(6): p. 503-20.
4. Kind, P., *The EuroQol instrument: an index of health-related quality of life*, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, Editor. 1996. p. 191-201.
5. Kaplan, R.M., et al., *The Quality of Well-being Scale. Applications in AIDS, cystic fibrosis, and arthritis.* Med Care, 1989. **27**(3 Suppl): p. S27-43.
6. Fitzpatrick, R., et al., *Evaluating patient-based outcome measures for use in clinical trials.* Health Technol Assess, 1998. **2**(14): p. i-iv, 1-74.
7. Rosenthal, R. and R.L. Rosnow, *Essentials of Behavioral Research: Methods and Data Analysis*. Second ed. 1991, New York: McGraw-Hill, Inc. 692.
8. Nunnally, J.C. and I.H. Bernstein, *Psychometric Theory*. Third ed. 1994, New York: McGraw-Hill, Inc. 752.
9. Elasy, T.A.M.D.M.P.H. and G.P. Gaddy, *Measuring Subjective Outcomes: Rethinking Reliability and Validity.* J Gen Intern Med, 1998. **13**(11): p. 757-761.
10. Messick, S., *Validity of psychological assessment: Validation of inferences from persons' responses to performances as scientific inquiry into score meaning.* American Psychologist, 1995. **50**(9): p. 741-749.

11. McHorney, C.A., et al., *The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups.* Med Care, 1994. **32**(1): p. 40-66.

12. Wu, A.W., et al., *Evidence for reliability, validity and usefulness of the Medical Outcomes Study HIV Health Survey (MOS-HIV).* Qual Life Res, 1997. **6**(6): p. 481-93.

13. Wu, A.W., et al., *A health status questionnaire using 30 items from the Medical Outcomes Study. Preliminary validation in persons with early HIV infection.* Med Care, 1991. **29**(8): p. 786-98.

14. Revicki, D.A., S. Sorensen, and A.W. Wu, *Reliability and validity of physical and mental health summary scores from the Medical Outcomes Study HIV Health Survey.* Med Care, 1998. **36**(2): p. 126-37.

15. Husted, J.A., et al., *Methods for assessing responsiveness: a critical review and recommendations.* J Clin Epidemiol, 2000. **53**(5): p. 459-68.

16. Bennett, K. and G. Torrance, *Measuring health preferences and utilities: rating scale, time trade-off and standard gamble methods*, in *Quality of Life and Pharmacoeconomics in Clinical Trials*, B. Spilker, Editor. 1996, Lippincott-Raven: Philadelphia. p. 235-265.

17. Lubeck, D.P., et al., *Quality of life and health service use among HIV-infected patients with chronic diarrhea.* J Acquir Immune Defic Syndr, 1993. **6**(5): p. 478-84.

18. Lubeck, D.P. and J.F. Fries, *Changes in quality of life among persons with HIV infection.* Qual Life Res, 1992. **1**(6): p. 359-66.

19. Lorenz, K.A., et al., *Associations of symptoms and health-related quality of life: findings from a national study of persons with HIV infection.* Ann Intern Med, 2001. **134**(9 Pt 2): p. 854-60.

20. Lubeck, D.P. and J.F. Fries, *Health status among persons infected with human immunodeficiency virus. A community-based study.* Med Care, 1993. **31**(3): p. 269-76.

21. Revicki, D.A., et al., *Quality of life outcomes of saquinavir, zalcitabine and combination saquinavir plus zalcitabine therapy for adults with advanced HIV infection with CD4 counts between 50 and 300 cells/mm3.* Antivir Ther, 1999. **4**(1): p. 35-44.

22. Revicki, D.A., A.W. Wu, and M.I. Murray, *Change in clinical status, health status, and health utility outcomes in HIV-infected patients.* Med Care, 1995. **33**(4 Suppl): p. AS173-82.

23. Lenert, L.A., A. Sturley, and M.E. Watson, *iMPACT3: Internet-based development and administration of utility elicitation protocols.* Med Decis Making, 2002. **22**(6): p. 464-74.

24. Lenert, L., et al., *Adverse effects of medication and trade-offs between length of life and quality of life in Human Immunodeficiency Virus infection.* Am J Med, 2002. **113**(3): p. 229-232.

25. Bult, J.R., et al., *Heterogeneity in the relationship between the time tradeoff and Short Form-36 for HIV-infected and primary care patients.* Med Care, 1998. **36**(4): p. 523-32.

26. Tsevat, J., et al., *Health values of patients infected with human immunodeficiency virus. Relationship to mental health and physical functioning.* Med Care, 1996. **34**(1): p. 44-57.

27. Cohen, J., *Statistical power analysis for the behavioral sciences*. 2nd ed. 1988, Hillsdale, New Jersey: Lawrence Erlbaum Associates Inc.

28. Miners, A.H., et al., *Health-related quality of life in individuals infected with HIV in the era of HAART.* HIV Clin Trials, 2001. **2**(6): p. 484-92.

29. Lamping, D.L., *Methods for measuring outcomes to evaluate interventions to improve health-related quality of life in HIV infection.* Psychology & Health, 1994. **9**(1-2): p. 31-49.

30. Miners, A.H., et al., *Assessing health-related quality-of-life in individuals with haemophilia.* Haemophilia, 1999. **5**(6): p. 378-85.

31. Copfer, A.E., et al., *The use of two measures of health-related quality of life in HIV-infected individuals: a cross-sectional comparison.* Qual Life Res, 1996. **5**(2): p. 281-6.

32. Karnofsky, D., et al., *The use of nitrogen mustards in the palliative treatment of carcinoma.* Cancer, 1948. **1**: p. 634-656.

33. Hughes, T.E., et al., *Construct validities of the Quality of Well-Being Scale and the MOS-HIV-34 Health Survey for HIV-infected patients.* Med Decis Making, 1997. **17**(4): p. 439-46.

34. Wu, A.W., et al., *Quality of life in a placebo-controlled trial of zidovudine in patients with AIDS and AIDS-related complex.* J Acquir Immune Defic Syndr, 1990. **3**(7): p. 683-90.

35. Kaplan, R.M., et al., *Validity of the Quality of Well-Being Scale for persons with human immunodeficiency virus infection. HNRC Group. HIV Neurobehavioral Research Center.* Psychosom Med, 1995. **57**(2): p. 138-47.

36. Anderson, J.P., et al., *Comparison of the Quality of Well-being Scale and the SF-36 results among two samples of ill adults: AIDS and other illnesses.* J Clin Epidemiol, 1998. **51**(9): p. 755-62.

37. Badia, X., et al., *A randomized study comparing instruments for measuring health-related quality of life in HIV-infected patients. Spanish MOS-HIV and MQOL-HIV Validation Group. Medical Outcomes Study HIV Health Survey.* Aids, 1999. **13**(13): p. 1727-35.

38. Barr, R.D., et al., *Health status and health-related quality of life associated with hemophilia.* Am J Hematol, 2002. **71**(3): p. 152-60.

39. Bayoumi, A.M. and D.A. Redelmeier, *Economic methods for measuring the quality of life associated with HIV infection.* Qual Life Res, 1999. **8**(6): p. 471-80.

40. Carr, A., et al., *A randomised, open-label comparison of three highly active antiretroviral therapy regimens including two nucleoside analogues and indinavir for previously untreated HIV-1 infection: the OzCombo1 study.* Aids, 2000. **14**(9): p. 1171-80.

41. Delate, T. and S.J. Coons, *The use of 2 health-related quality-of-life measures in a sample of persons infected with human immunodeficiency virus.* Clin Infect Dis, 2001. **32**(3): p. E47-52.

42. Djulbegovic, B., et al., *Comparison of the quality of life between HIV positive hemophilia patients and HIV negative hemophilia patients.* Haemophilia, 1996. **2**: p. 166-172.

43. Johnson, E.S., et al., *A utility assessment of oral and intravenous ganciclovir for the maintenance treatment of AIDS-related cytomegalovirus retinitis.* Pharmacoeconomics, 1996. **10**(6): p. 623-9.

44. Kempen, J.H., et al., *The effect of cytomegalovirus retinitis on the quality of life of patients with AIDS in the era of highly active antiretroviral therapy.* Ophthalmology, 2003. **110**(5): p. 987-95.

45. Mrus, J.M., et al., *Utilities derived from visual analog scale scores in patients with HIV/AIDS.* Med Decis Making, 2003. **23**(5): p. 414-21.

46. Nickel, J.T., et al., *Quality of life in nurse case management of persons with AIDS receiving home care.* Res Nurs Health, 1996. **19**(2): p. 91-9.

47. Group, N.M.R.C.H.T., *Double blind dose-response study of zidovudine in AIDS and advanced HIV infection. Nordic Medical Research Councils' HIV Therapy Group.* Bmj, 1992. **304**(6818): p. 13-7.

48. Oliva, J., C. Roa, and J. del Llano, *Indirect costs in ambulatory patients with HIV/AIDS in Spain: a pilot study.* Pharmacoeconomics, 2003. **21**(15): p. 1113-21.

49. Revicki, D.A., et al., *Quality of life outcomes of combination zalcitabine-zidovudine, saquinavir-zidovudine, and saquinavir-zalcitabine-zidovudine therapy for HIV-infected adults with CD4 cell counts between 50 and 350 per cubic millimeter. PISCES (SV14604) Study Group.* Aids, 1999. **13**(7): p. 851-8.

50. Schag, C.A., et al., *Assessing the needs and quality of life of patients with HIV infection: development of the HIV Overview of Problems-Evaluation System (HOPES).* Qual Life Res, 1992. **1**(6): p. 397-413.

51. Tramarin, A., et al., *An economic evaluation of home-care assistance for AIDS patients: a pilot study in a town in northern Italy.* Aids, 1992. **6**(11): p. 1377-83.

52. Trippoli, S., et al., *Multivariate analysis of factors influencing quality of life and utility in patients with haemophilia.* Haematologica, 2001. **86**(7): p. 722-8.

53. Tsevat, J., et al., *The will to live among HIV-infected patients.* Ann Intern Med, 1999. **131**(3): p. 194-8.

54. Wu, A.W., et al., *The effect of mode of administration on medical outcomes study health ratings and EuroQol scores in AIDS.* Qual Life Res, 1997. **6**(1): p. 3-10.

55. Wu, A.W., et al., *Validity and responsiveness of the euroqol as a measure of health-related quality of life in people enrolled in an AIDS clinical trial.* Qual Life Res, 2002. **11**(3): p. 273-82.

56. Zinkernagel, C., et al., *Importance of mental health assessment in HIV-infected outpatients.* J Acquir Immune Defic Syndr, 2001. **28**(3): p. 240-9.

57. Turner, B.J., J.V. Kelly, and J.K. Ball, *A severity classification system for AIDS hospitalizations.* Med Care, 1989. **27**(4): p. 423-37.